# Section 11: Exploring Bivariate Data

The following maps the videos in this section to the Texas Essential Knowledge and Skills for Mathematics TAC §111.47(c).

**11.01 Scatterplots**

- Statistics (7)(A)

**11.02 Correlation**

- Statistics (7)(A)

**11.03 Determining the Line of Best Fit**

- Statistics (7)(A)
- Statistics (7)(B)
- Statistics (7)(C)

**11.04 Making Predictions**

- Statistics (7)(F)

**11.05 Interpreting Slope and y-intercept**

- Statistics (7)(F)

**11.06 The Median-Median Line and Least Absolute Value Line**

- Statistics (7)(A)
- Statistics (7)(C)
- Statistics (7)(D)

**11.07 Outliers and Influential Points**

- Statistics (7)(A)
- Statistics (7)(E)

# 11.01
# Scatterplots

A ***scatterplot*** is a graphical representation of the relationship between two quantitative variables.

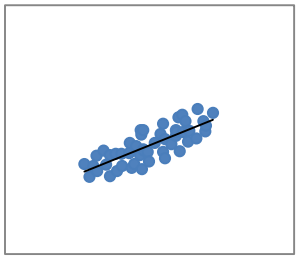
$x$
- _____ variable
- _____ variable

$y$
- _____ variable
- _____ variable

We use scatterplots to visually assess the strength, direction, and overall pattern of the relationship.

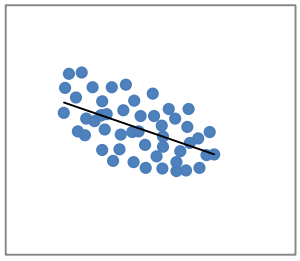
- ***Strength*** – Strong or weak
- ***Direction*** – Positive (as $x \uparrow, y \uparrow$) or negative (as $x \uparrow, y \downarrow$)
- ***Overall pattern*** – Linear, nonlinear, or no pattern
- ***Outliers*** – Data points that don't follow the overall pattern

Strong, positive, linear



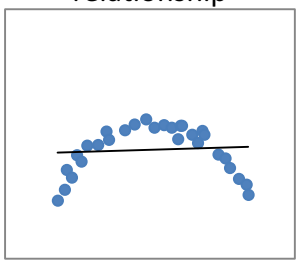
As $x$ increases, $y$ increases.

Moderate, negative, linear



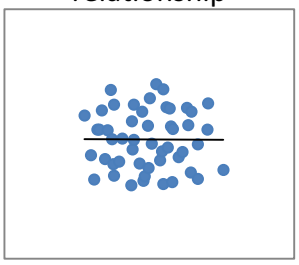
As $x$ increases, $y$ decreases.

Nonlinear relationship



The rate of change is not constant.

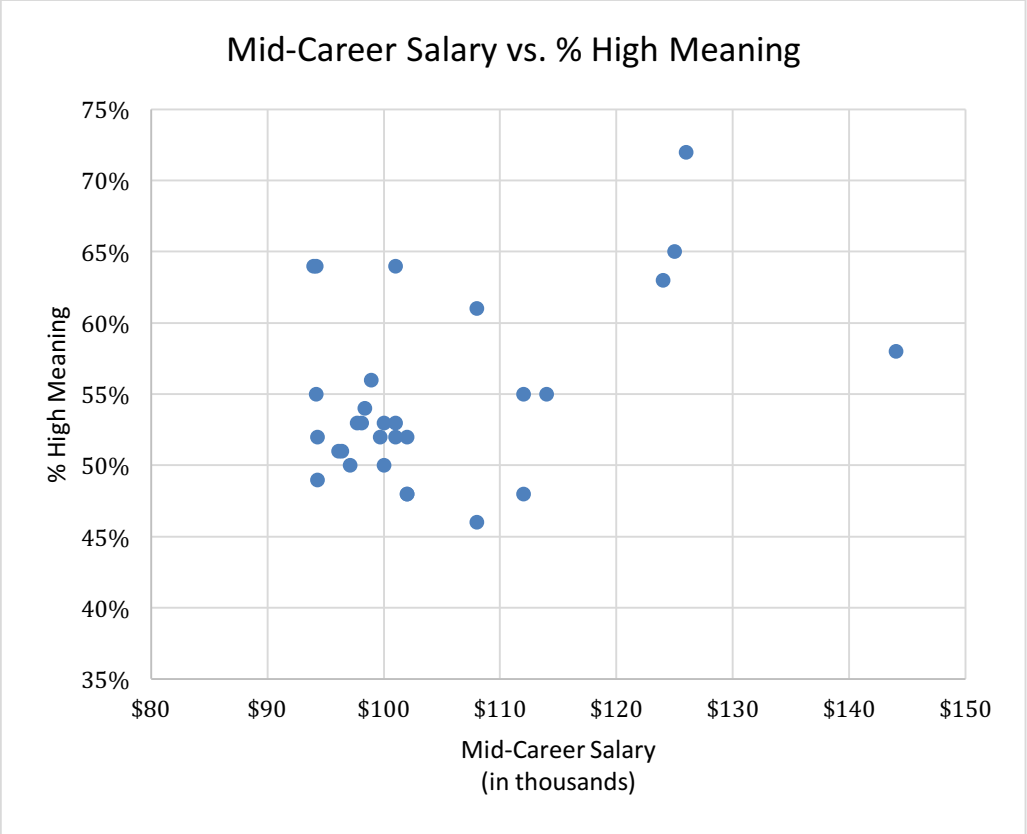
No relationship



There is no relationship between $x$ and $y$.

The table below lists the top 30 public universities based on the median mid-career salary of alumni who earned a bachelor's degree but not an advanced degree. The table also shows the percentage of bachelor's alumni with high-meaning careers—that is, alumni who say their work makes the world a better place (PayScale, n.d.; U.S. News, n.d.).

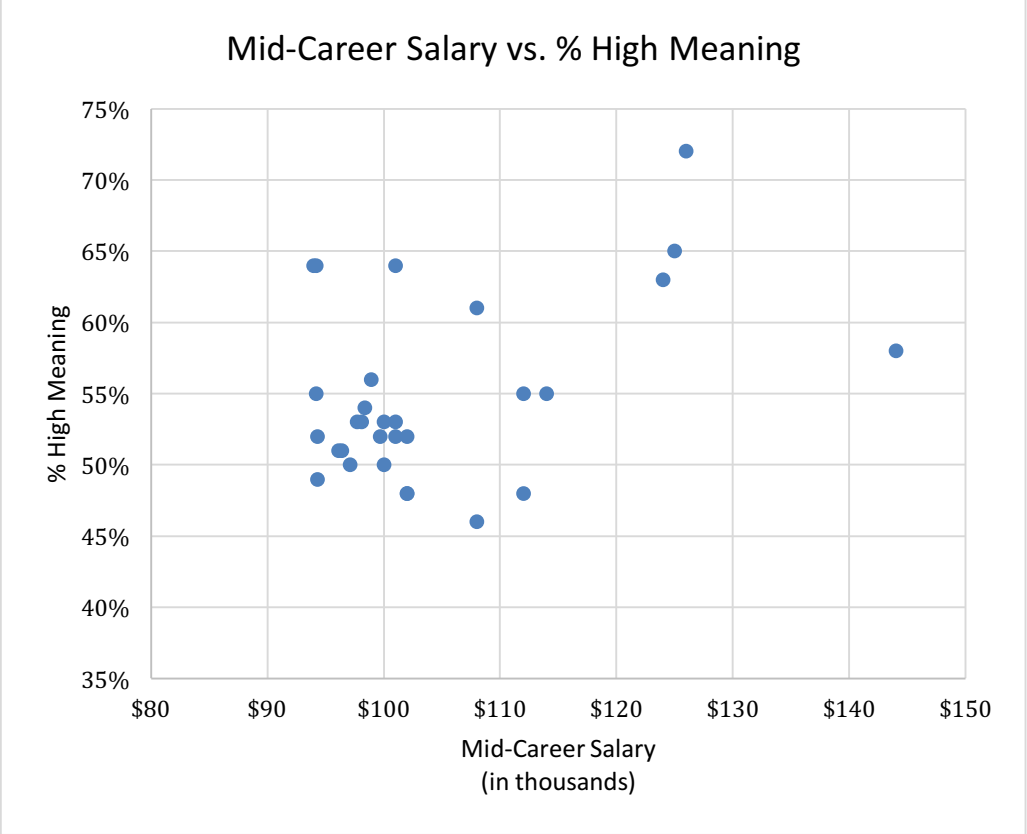
| School | Mid-Career Salary (in thousands) | % High Meaning |
|---|---|---|
| SUNY - Maritime College | $144.0 | 58% |
| United States Military Academy | $126.0 | 72% |
| United States Naval Academy | $125.0 | 65% |
| United States Air Force Academy | $124.0 | 63% |
| Colorado School of Mines | $114.0 | 55% |
| Georgia Institute of Technology | $112.0 | 48% |
| University of California - Berkeley | $112.0 | 55% |
| University of California - San Diego (UCSD) | $108.0 | 61% |
| Virginia Military Institute | $108.0 | 46% |
| California Polytechnic State University (CalPoly) | $102.0 | 52% |
| Michigan Technological University | $102.0 | 48% |
| University of Illinois at Urbana-Champaign (UIUC) | $102.0 | 48% |
| Missouri University of Science and Technology (S&T) | $101.0 | 53% |
| South Dakota School of Mines & Technology | $101.0 | 64% |
| University of California - Santa Barbara (UCSB) | $101.0 | 52% |
| Massachusetts Maritime Academy | $100.0 | 50% |
| New Jersey Institute of Technology (NJIT) | $100.0 | 53% |
| Texas A&M University | $99.7 | 52% |
| University of Michigan - Ann Arbor | $98.9 | 56% |
| University of California - Davis (UC Davis) | $98.4 | 54% |
| University of California - Los Angeles (UCLA) | $98.1 | 53% |
| San Jose State University (SJSU) | $97.7 | 53% |
| University of Virginia (UVA) - Main Campus | $97.1 | 50% |
| University of Texas (UT) - Austin | $96.4 | 51% |
| Virginia Tech | $96.1 | 51% |
| Stony Brook University | $94.3 | 52% |
| University of Colorado - Boulder (CU) | $94.3 | 49% |
| Purdue University - Main Campus | $94.2 | 55% |
| University of Washington (UW) | $94.2 | 64% |
| University of Alabama | $94.0 | 64% |




1. Describe the relationship between mid-career salary and the percentage of alumni who say their work makes the world a better place.

# 11.02
# Correlation

The scatterplot below shows the relationship between mid-career salary and high-meaning career percentage for the top 30 public universities based on mid-career salary (PayScale, n.d.; U.S. News, n.d.). Describe the relationship.



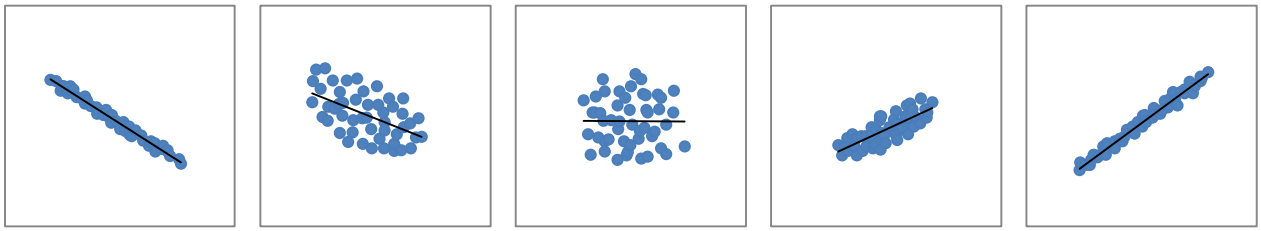

The ***correlation coefficient***, $r$, measures the strength and direction of the linear association between two quantitative variables.

- $r = \frac{1}{n-1}\sum z_x z_y = \frac{1}{n-1}\sum\left(\frac{x-\bar{x}}{s_x}\right)\left(\frac{y-\bar{y}}{s_y}\right)$
- $-1 \leq r \leq +1$
- The correlation coefficient ($r$) has no units.

1. Using the values in the boxes, indicate which of the following values of $r$ best describes each of the scatterplots.

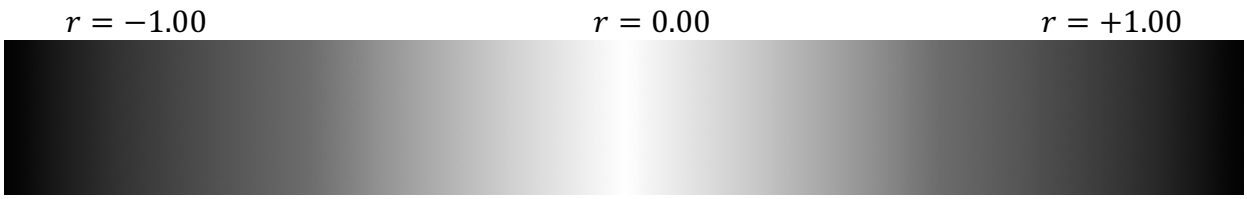
| | | | | |
|---|---|---|---|---|
| $r = -0.001$ | $r = +0.790$ | $r = -0.991$ | $r = +0.990$ | $r = -0.547$ |

$r =$ _____ $r =$_____ $r =$_____ $r =$_____ $r =$_____



- The closer the points are to the line, the _____ the absolute value of $r$ will be.
- The closer $r$ is to zero, the _____ the linear relationship is between $x$ and $y$.
- The values $r = +0.450$ and $r = -0.450$ both indicate the _____ strength of association between the variables.
- The correlation coefficient ($r$) is not affected by units.
- The correlation coefficient does not change if we switch $x$ and $y$.
- The correlation coefficient is strongly affected by outliers.

**Strength of a Linear Relationship**




Perfect Negative Linear Relationship

No Linear Relationship

Perfect Positive Linear Relationship

STUDY EDGE TIP

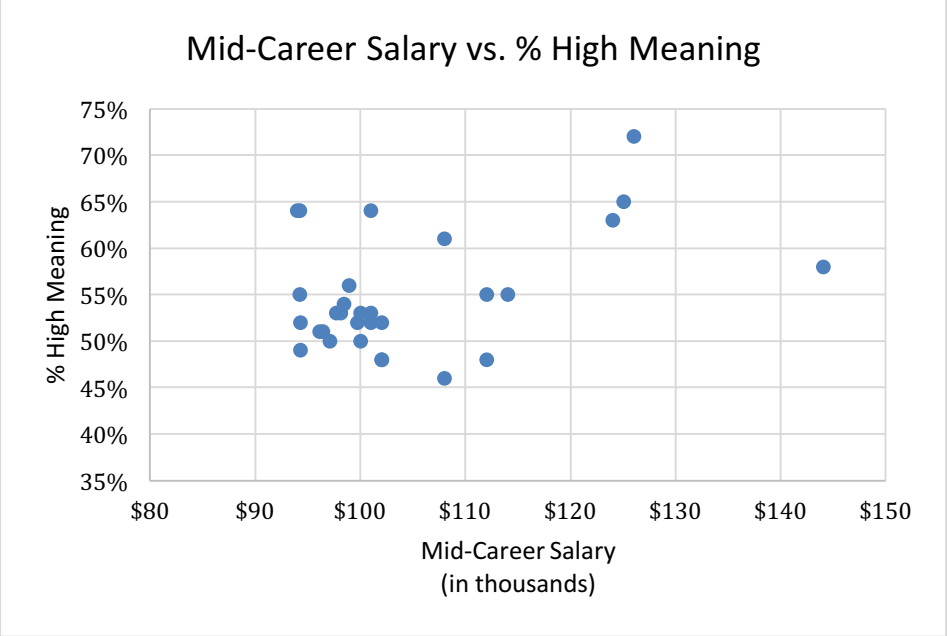
Correlation does not imply causation.

2. The table below lists the top 30 public universities based on the median mid-career salary of alumni who earned a bachelor's degree but not an advanced degree. The table also shows the percentage of bachelor's alumni with high-meaning careers—that is, alumni who say their work makes the world a better place (PayScale, n.d.; U.S. News, n.d.). Calculate and interpret the correlation coefficient. Access the data at tiny.cc/se-correlation.

| School | Mid-Career Salary (in thousands) | % High Meaning |
|---|---|---|
| SUNY - Maritime College | $144.0 | 58% |
| United States Military Academy | $126.0 | 72% |
| United States Naval Academy | $125.0 | 65% |
| United States Air Force Academy | $124.0 | 63% |
| Colorado School of Mines | $114.0 | 55% |
| Georgia Institute of Technology | $112.0 | 48% |
| University of California - Berkeley | $112.0 | 55% |
| University of California - San Diego (UCSD) | $108.0 | 61% |
| Virginia Military Institute | $108.0 | 46% |
| California Polytechnic State University (CalPoly) | $102.0 | 52% |
| Michigan Technological University | $102.0 | 48% |
| University of Illinois at Urbana-Champaign (UIUC) | $102.0 | 48% |
| Missouri University of Science and Technology (S&T) | $101.0 | 53% |
| South Dakota School of Mines & Technology | $101.0 | 64% |
| University of California - Santa Barbara (UCSB) | $101.0 | 52% |
| Massachusetts Maritime Academy | $100.0 | 50% |
| New Jersey Institute of Technology (NJIT) | $100.0 | 53% |
| Texas A&M University | $99.7 | 52% |
| University of Michigan - Ann Arbor | $98.9 | 56% |
| University of California - Davis (UC Davis) | $98.4 | 54% |
| University of California - Los Angeles (UCLA) | $98.1 | 53% |
| San Jose State University (SJSU) | $97.7 | 53% |
| University of Virginia (UVA) - Main Campus | $97.1 | 50% |
| University of Texas (UT) - Austin | $96.4 | 51% |
| Virginia Tech | $96.1 | 51% |
| Stony Brook University | $94.3 | 52% |
| University of Colorado - Boulder (CU) | $94.3 | 49% |
| Purdue University - Main Campus | $94.2 | 55% |
| University of Washington (UW) | $94.2 | 64% |
| University of Alabama | $94.0 | 64% |

# 11.03
# Determining the Line of Best Fit

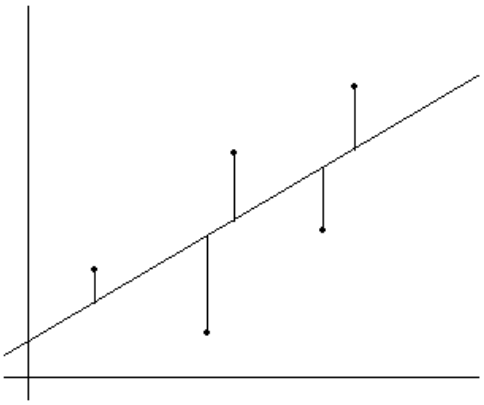
The scatterplot below describes the relationship between mid-career salary and high-meaning career percentage for the top 30 public universities based on mid-career salary (PayScale, n.d.; U.S. News, n.d.).




What function best fits this data? Draw the function on the scatterplot.

There are several ways to determine the line that best fits the data, or the ***line of best fit***. The most widely used is the ***least-squares regression line***.

**Residuals**



The residuals (_____)

- _____ distance between the $y$-value and the line
- Residual = actual $y$ – predicted $y$
  - $e = y - \hat{y}$
- Actual $y > \hat{y} \rightarrow e = (+)$
- Actual $y < \hat{y} \rightarrow e = (-)$
- Actual $y = \hat{y} \rightarrow e = 0$
- Residuals always sum to ____ and $\bar{e} =$ ____.

**Least-Squares Regression Line**

- The least-squares regression line always goes through the point _____.
- The least-squares regression line is affected by outliers.
- The least-squares regression line _____ the sum of the residuals _____ → minimizes $\sum(y - \hat{y})^2$.
- In algebra, $y = mx + b$.
- In statistics, $\hat{y} = a + bx$, where
  - $b = \text{slope} = r\left(\frac{s_y}{s_x}\right)$ and
  - $a = \ y - \text{intercept} = \bar{y} - b\bar{x}$.

Consider the two lines that pass through $(\bar{x}, \bar{y})$.

Line #1

| $e$ | $e^2$ |
|---|---|
| -2 | 4 |
| 1 | 1 |
| 1 | 1 |
| 0 | 6 |

Line #2

| $e$ | $e^2$ |
|---|---|
| 1 | 1 |
| 0 | 0 |
| -1 | 1 |
| 0 | 2 |

Line #2 is better, because it has a *smaller* sum of residuals squared.

1. Use the information below to determine and graph the line of best fit used to predict a university's percentage of alumni who think their work makes the world a better place based on the median mid-career salary of the university's alumni.

| School | Mid-Career Salary (in thousands) | % High Meaning |
|---|---|---|
| SUNY - Maritime College | \$144.0 | 58% |
| United States Military Academy | \$126.0 | 72% |
| ⋮ | ⋮ | ⋮ |
| University of Washington (UW) | \$94.2 | 64% |
| University of Alabama | \$94.0 | 64% |

| | |
|---|---|
| $\bar{x} = \$104.5$ | $\bar{y} = 54.9\%$ |
| $s_x = \$11.8$ | $s_y = 6.3\%$ |
| $r = 0.4147$ | |

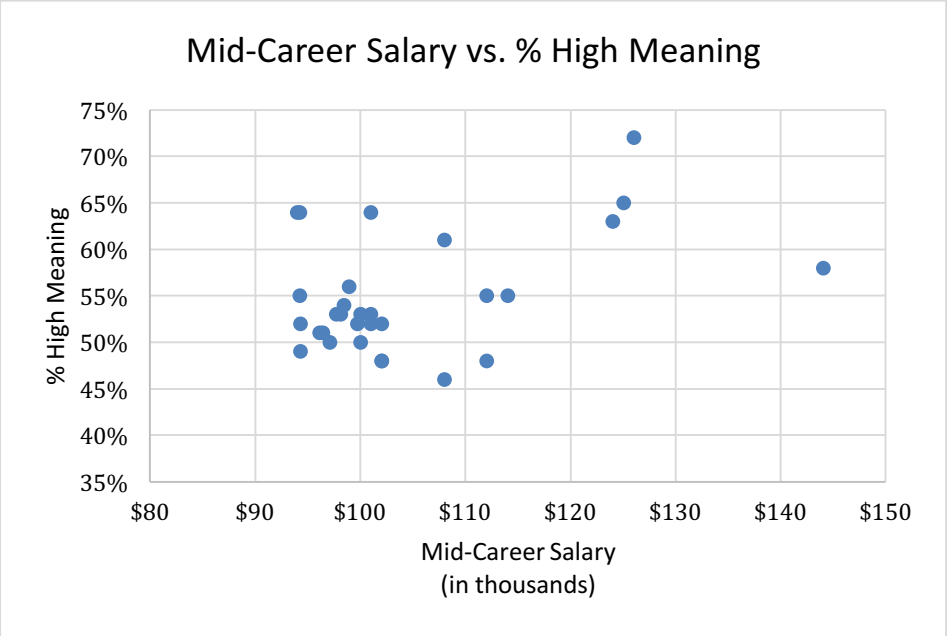
STUDY EDGE TIP

Always make sure the sign on $r$ matches the sign of the slope, $b$.

# 11.04
# Making Predictions

The line of best fit can be used to model the relationship between two variables, make predictions for $y$ given a value of $x$, and understand how changes in $x$ affect $y$.

The least-squares regression equation $\hat{y} = 31.89 + 0.22x$ estimates the relationship between $y =$ percentage of alumni who think their work makes the world a better place and $x =$ median mid-career salary (in thousands) for the top 30 public universities based on mid-career salary.




1. Use the equation to predict the percentage of alumni who think their work makes the world a better place for a university with a mid-career salary of $94,000.

2. Use the equation to predict the percentage of alumni who think their work makes the world a better place for a university with a mid-career salary of $294,000.

3. University of Alabama alumni have a median mid-career salary of $94,000, and 64% of alumni say they have high-meaning jobs. Calculate the residual for the University of Alabama.

# 11.05
# Interpreting Slope and $y$-Intercept

**Interpreting Slope**

- $b = \text{slope} = r\left(\frac{s_y}{s_x}\right)$
- The slope tells us how $y$ changes with respect to $x$.
- Slope is expressed in $y$-units per $x$-unit.
- Slope tells us how the $y$ variable changes for a one-unit increase in the $x$ variable.

**Interpreting the $y$-Intercept**

- $a = \bar{y} - b\bar{x}$
- The $y$-intercept tells us where the line intercepts the $y$-axis.
- Be cautious when interpreting the $y$-intercept.

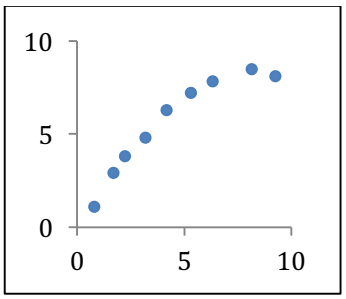
The least-squares regression equation $\hat{y} = 31.89 + 0.22x$ estimates the relationship between $y =$ percentage of alumni who think their work makes the world a better place and $x =$ median mid-career salary (in thousands) for the top 30 public universities based on mid-career salary.

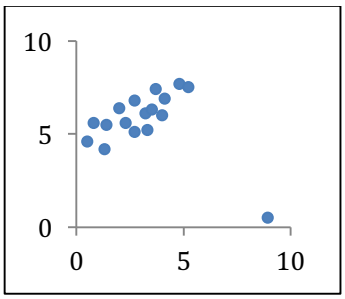
1. Interpret the slope.

2. Interpret the $y$-intercept.

# 11.06
# The Median–Median Line and Least Absolute Value Line

Is the least-squares regression line appropriate for estimating the relationships below? Why or why not?



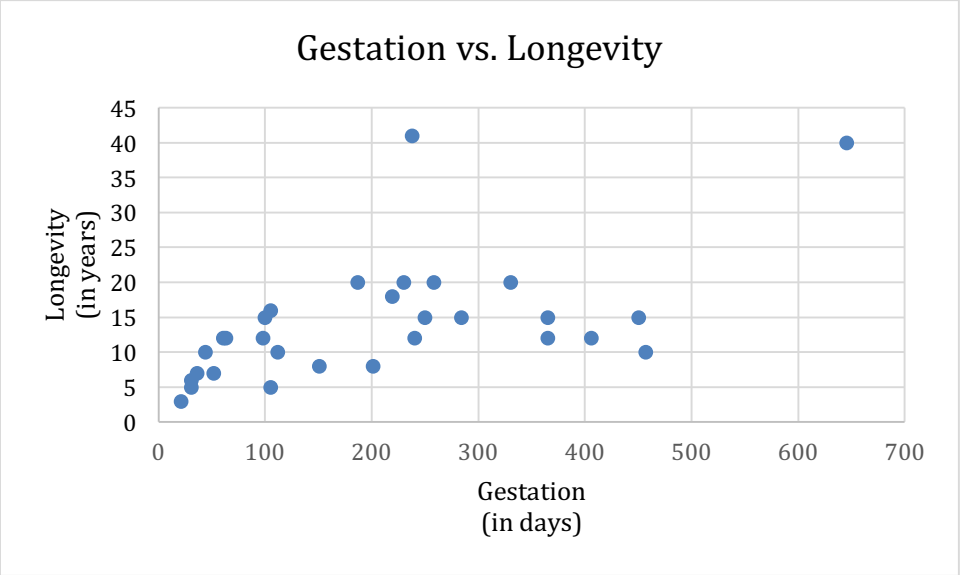

The ***least absolute value line***, sometimes referred to as ***L1 regression***, is an alternative method used to estimate the trend line. Rather than minimizing the sum of the *squared* errors, it minimizes the sum of the *absolute* error. Why would this method be useful when there are outliers present?

The ***median–median line*** is a less common method of fitting a line to data than the least-squares method, but it is useful when there are outliers away from the overall pattern in the scatterplot.

To fit the median–median line:

1) Sort the data by the independent variable.
2) Divide the sorted data into three groups.
3) Find the median $(x, y)$ of each group.
4) Find the slope of the line that passes through $(x_1, y_1)$ and $(x_3, y_3)$.
$$b = \frac{y_3 - y_1}{x_3 - x_1}$$
5) Use each median point to find the $y$-intercept.
$$a = \frac{(y_1 - bx_1) + (y_2 - bx_2) + (y_3 - bx_3)}{3}$$

The following data and scatterplot summarize the average gestation period, in days, and average longevity, in years, for a sample of 30 animals as reported in *The World Almanac and Book of Facts* (McGeveran, 2006).




| Animal | Gestation (in days) | Longevity (in years) | Animal | Gestation (in days) | Longevity (in years) | Animal | Gestation (in days) | Longevity (in years) |
|---|---|---|---|---|---|---|---|---|
| Mouse | 21 | 3 | Beaver | 105 | 5 | Elk | 250 | 15 |
| Chipmunk | 31 | 6 | Tiger | 105 | 16 | Gorilla | 258 | 20 |
| Rabbit | 31 | 5 | Pig | 112 | 10 | Cow | 284 | 15 |
| Kangaroo | 36 | 7 | Goat | 151 | 8 | Horse | 330 | 20 |
| Squirrel | 44 | 10 | Baboon | 187 | 20 | Donkey | 365 | 12 |
| Fox | 52 | 7 | Deer | 201 | 8 | Zebra | 365 | 15 |
| Dog | 61 | 12 | Black Bear | 219 | 18 | Camel | 406 | 12 |
| Cat | 63 | 12 | Chimpanzee | 230 | 20 | Rhinoceros | 450 | 15 |
| Leopard | 98 | 12 | Hippopotamus | 238 | 41 | Giraffe | 457 | 10 |
| Lion | 100 | 15 | Moose | 240 | 12 | Elephant | 645 | 40 |

1. Are there any outliers? If so, how would they affect the least-squares regression line?

2. Find and graph the median–median line.

   i. Sort the data by the independent variable.

   ii. Divide the sorted data into three groups.

   iii. Find the median $(x, y)$ of each group.

   iv. Find the slope of the line that passes through $(x_1, y_1)$ and $(x_3, y_3)$.

$$b = \frac{y_3 - y_1}{x_3 - x_1}$$

   v. Use each median point to find the $y$-intercept.

$$a = \frac{(y_1 - bx_1) + (y_2 - bx_2) + (y_3 - bx_3)}{3}$$

   vi. Write and graph the median–median line.

3. The least-squares regression line for the data is $\hat{y} = 7.6 + 0.03x$. Draw the line on the scatterplot.

4. Which line appears to be a better fit—the least-squares regression line or the median-median line?

# 11.07
# Outliers and Influential Points

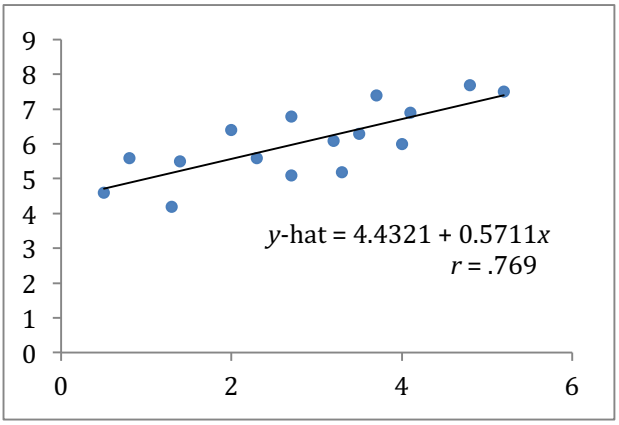
***Outliers*** are unusual observations that do not follow the overall pattern of the data. Outliers can be extreme points in the $x$and/or $y$ direction.

An outlier is an ***influential point*** if it has a large effect on the slope of the regression equation.

Outliers and influential points can be identified by

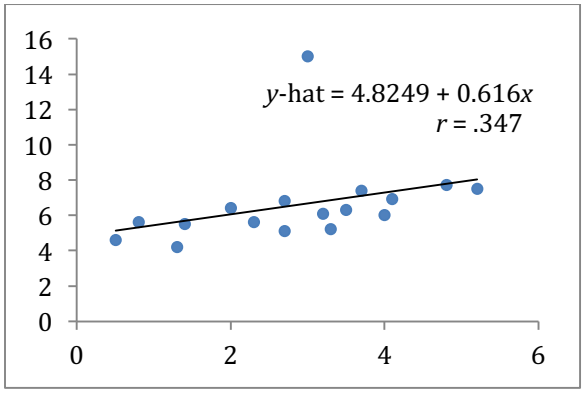
- examining residuals,
- looking at scatterplots, and
- using technology.

Consider the scatterplot below that contains no outliers.

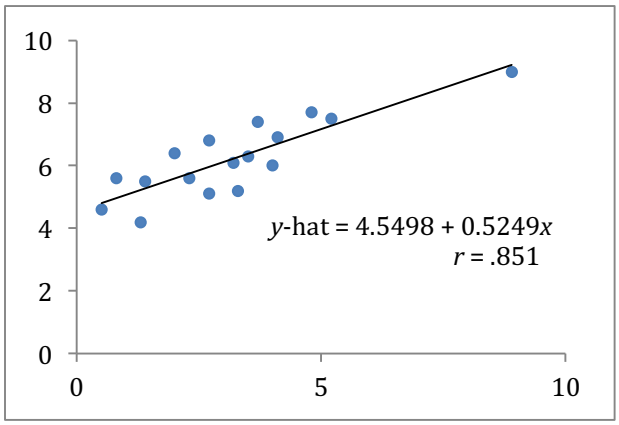



1. For each of the following, circle the outlier, identify the direction of the outlier, and identify the repercussions of the outlier when it is included in the data set.

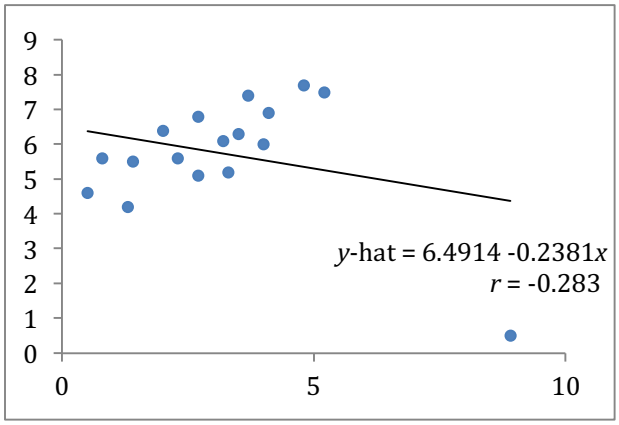
   i.



ii.




iii.




2. Use the dataset that includes the influential point, which can be found at tiny.cc/se-outliers, to explore the effects of influential points and outliers. Draw conclusions based on your findings.

## References

McGeveran, W. A. (2006). *The World Almanac and Book of Facts: 2006*. New York: World Almanac Books.

PayScale. (n.d.). "Best Public Universities by Salary Potential." Retrieved May 2017 from https://www.payscale.com/college-salary-report/best-schools-by-type/bachelors/public-schools

U.S. News & World Report. (n.d.). "Top Public National Universities." Retrieved May 2017 from https://www.usnews.com/best-colleges/rankings/national-universities/top-public